Web2Web : Publicisation sur le Web
des Collections Événementielles et Médiatiques Enrichies issues du web 2.0.

Le projet que nous déposons vise à définir un cadre propice à la constitution et à l'animation d'un groupe de travail thématique susceptible, à une échelle locale et dans un premier temps, d'agréger une expertise et une expérience nécessaire à la mise en ligne d'une collection réalisée dans le cadre d'un financement ANR 2016-2017. L'ambition est de tirer avantage de ce cas d'école pour installer dans la durée une réflexion interdisciplinaire se rapportant à la publicisation de collections numériques issues des réseaux sociaux dans un contexte de *big data*.
Cette réflexion fait écho aux questionnements développés dans différents domaines des Sciences Humaines et Sociales portant sur la disponibilité de dispositifs documentaires et de documents nouveaux ainsi que sur le renouvellement des modalités et des pratiques scientifiques expérimentales et empiriques. L'objectif est ainsi de s'inscrire dans une éthique de l'accessibilité et de l'ouverture des données de la recherche et, plus avant, dans la promotion d'un accès public aux données du web dans le sens de *l'open data*.

PARTENAIRES DU PROJET :

- **Institutions :**
  - la Bibliothèque Nationale de France (BnF) – Collections numériques - dépôt légal
  - CNIL
- **Réseaux scientifiques :**
  - Consortium IIPC : International Internet Preservation Consortium,
  - Réseau mathe-shs : réseau CNRS des professionnels de l'analyse de données du web,
  - GIS IPAPIC : Institutions patrimoniales et pratiques interculturelles.
  - Consortium ANR RSJ-MéDiS
- **Laboratoires associés :** PACTE, GRESEC, LIDILEM

CONTEXTE DU PROJET :

Depuis 2010, à l'occasion de projet de recherche se rapportant à la médiatisation des événements grand public, politiques ou sportifs (JO2010, JO2012, JO2014, Présidentielles 2012, Européennes 2014), le laboratoire PACTE a réalisé des collections volumineuses d'informations et de documents numériques extraits du web, en liens avec la capture de flux de messages publiés sur des réseaux sociaux et Twitter en particulier. Chacune de ces collections couvre de 3 à 9 mois d'enregistrements continus se rapportant à plusieurs millions de Tweets. Elles comportent également des métadonnées et des enrichissements issus de traitements linguistiques[1] et statistiques, visant à en faciliter l'analyse et l'interprétation scientifique. Depuis 2012, l'enrichissement comporte également des références et des métadonnées fournies par la bibliothèque nationale de France (BnF) dans le cadre d'une convention avec le laboratoire PACTE. Ces informations sont issues des dispositifs d'archivage mis en œuvre par la BnF dans le cadre de sa mission de préservation du web. Elles constituent un pont entre collections et ouvrent la voie à de nouvelles modalités exploratoires des archives du web dont l'usage était jusqu'alors très limité.
Parallèlement à cette démarche d'archivage, le laboratoire PACTE a entrepris une réflexion avec la CNIL portant sur les modalités de constitution, d'exploitation et de valorisation de ces archives jusque-là informelles. La nature des problèmes soulevés, notamment au regard de contenus multimédia, ne permet pas d'apporter en l'état, une réponse immédiate et satisfaisante à la fois du point de vue légal et du point de vue scientifique. Ces échanges soulignent l'importance d'un travail de fond susceptible de constituer, comme avec la BnF, un contexte d'innovation essentiel.
Dans le même temps une action de recherche « Humanités et Numérique » (GRESEC) questionne les conséquences en termes de représentations intellectuelle et visuelle de la constitution de fonds issus de différents terrains d'analyse. Elle souligne la nécessité d'une réflexion éthique sur les choix effectués.

[1] Services de traitement de la langue (TAL) développés par le LIDILEM dans le cadre de projets communs, notamment un projet PEPS Humain (2014-2015).

OBJECTIFS DU PROJET:

Le laboratoire PACTE est désormais impliqué dans l'ANR RSJ-MéDiS[2] (2016-2017) à titre de porteur d'un volet d'étude des productions médiatiques associées aux grands événements sportifs (JO2016). Dans cette perspective de nouveaux enrichissements ont été apportés à la constitution de collections : fils d'actualités médiatiques, flux iconographiques. Ces enrichissements découlent des acquis des précédents projets et des collaborations qui se sont nouées entre les laboratoires PACTE et LIDILEM qui est pour sa part également investi dans la TGIR[3] Huma-Num..

La proximité que le projet RSJ-MéDiS introduit, avec l'un des acteurs clé (FranceTélévisions) de la mise en discours de l'événement et notamment des sources de publications qu'il filtre et traduit, devrait permettre d'étoffer l'enrichissement et lui conférer une valeur référentielle importante au sein des SHS et des SIC en particulier.

Dans une démarche expérimentale et empirique, nous souhaitons désormais constituer et développer une réflexion privilégiant deux dimensions :

- **Juridique et déontologique :**
  - La capture de documents dans la perspective de constitution de corpus interroge la propriété intellectuelle. A l'origine, des sources, des publications, des personnes revendiquent la qualité d'auteur ayant des droits à faire valoir sur leurs créations dès lors qu'ils estiment faire œuvre originale. La question de la reconnaissance d'une « paternité intellectuelle » ne saurait être éludée. De même, les archives du web peuvent revêtir la forme d'images relatives à des personnes, des groupes de personnes ou des biens. La confrontation entre des droits de la personne par exemple et le droit à l'information se trouve posé, avec d'autant plus de pertinence que se profile le concept de consentement des personnes.
  - Si ces premières questions trouvent en partie des réponses juridiques, il nous appartient également de mettre en place en amont une démarche déontologique, garante des droits des personnes mais aussi de la pertinence de la constitution d'archives.
  - Cette réflexion impose une double démarche : un état des lieux de la réglementation et la mise en place d'une régulation consentie entre les professionnels de la constitution d'archives du web et les personnes concernées, qu'elles soient physiques ou morales. Le projet de loi République Numérique en cours de discussion, ouvre à cet égard des perspectives stimulantes.

- **Documentaire, patrimoniale et éthique :**

  La légitimité de ce qui est archivé et potentiellement versé au patrimoine impose également une réflexion éthique sur la constitution de collections documentaires stables à partir d'une matière première instable, fortement contextualisée et parfois perçue par les producteurs comme éphémère (conversations). Dans ce cadre nous proposons d'établir des préconisations, d'évaluer des méthodes et des procédures pour les points suivants :
  - Les choix de la structure de données ainsi que des formats (normés ou non) de représentation dans les archives dépendent des contraintes liées aux problématiques juridiques et légales (anonymisation, etc.) et aux impératifs de conservation et de diffusion ;
  - Les processus d'enrichissement et d'élaboration de métadonnées sont impliqués à deux niveaux: d'une part, celui de la caractérisation des archives, de son élaboration et des modalités d'usage de ses contenus ; d'autre part, celui de ses contenus et des modalités de sélection permettant la constitution de corpus. Intégrer ces différents niveaux de description dans un processus documentaire cohérent, supportant un principe d'interopérabilité avec l'archivage du web notamment ;
  - Enfin, les traitements linguistiques (tagging lexical, segmentation, repérage d'entités nommées, etc.) qui sont déjà incorporés à la structure de données des enregistrements (ou susceptibles de l'être en post traitement) contribuent à la valorisation de l'archive. L'augmentation des données nécessite de

[2] ANR Responsabilité Sociale des Journalistes : Médias, Diversité et Sport en partenariat avec les laboratoires : CRAPE (UMR 6051), Praxiling (UMR 5267 CNRS), Geriico (EA 4073) et l'URePSSS (EA 4110) et avec le réseau France Télévision (FTV)

[3] Très Grande Infrastructure Numérique (TGIR): http://www.huma-num.fr/.

travailler avec les spécialistes du TAL et de l'analyse linguistique pour adapter les formats aux environnements d'analyse et aux pratiques de la recherche dans ces domaines.

Le groupe de travail est constitué des contributeurs initiaux du projet auquel pourront s'agréger des chercheurs (doctorants, enseignants-chercheurs) du site au fur et à mesure des avancées, des problèmes rencontrés ou des manifestations d'intérêt. L'objectif est ainsi de faire émerger une cartographie et un réseau de compétences de site. Pour mener à bien ce projet de nature collaborative et interdisciplinaire, nous nous proposons d'utiliser les outils de la plate-forme PAMPA pour animer le groupe de travail et rendre visible les avancées de nos réflexions. Nous souhaitons également mettre à profit les dynamiques thématiques notamment avec les projets de l'axe Corpus hébergés par la MSHAlpes et/ou développés dans le réseau des MS.

À l'issue d'une année du projet, nous espérons être à même de :

- **Mettre en ligne la collection** produite à l'occasion des Jeux Olympiques d'été 2016 (Rio), ayant les caractéristiques techniques et juridiques nécessaires à leur diffusion et leur appropriation dans un contexte scientifique ouvert.
- **Établir un guide de préconisations** techniques et méthodologiques reprenant les 3 dimensions précédentes.

CALENDRIER

- janvier 2016 : 1ère réunion du groupe de travail (suivantes : avril, septembre, novembre)
- février-mai : établissement du cahier des charges de mise en ligne
- juin 2016 : séminaire 1
- juin-août : mise en œuvre des préconisations dans les formats et structure de données
- septembre-novembre : adaptations documentaires des collections (+IE)
- décembre 2016 : séminaire 2+ publication des collections

Séminaire 1 : thématique envisagée : éléments techniques de l'*open acces* et de l'*open data*, contraintes de mise en ligne et de publicisation des données du web.
Séminaire 2 : restitution des travaux, présentation des éléments méthodologiques.

Projet de publication sur Ortolang (https://www.ortolang.fr) qui est l'un des services proposés par la TGIR Huma-Num.

BUDGET : 14 500€

	Qte	Individus	nuités	Montant
Déplacement France (Paris)	2	4	1	3044
Vacation IE	3	1		6000
Déplacement colloque	1	3	3	2730
Traduction	1			1700
Réunion groupe travail	4	10		800
Séminaire	2	20		226
			Total	14500

- **PORTEUR DU PROJET :** Jean-Marc Francony (PACTE) : jeanmarc.francony@umrpacte.fr

CONTRIBUTEURS ET MEMBRES DU GROUPE DE TRAVAIL :
Anne-Marie Benoit (PACTE), Jean-Stéphane Carnel (GRESEC), Jean-Marc Francony (PACTE), Aude Inaudi (GRESEC), Françoise Papa (PACTE), Claude Ponton (LIDILEM), Annie-Claude Salomon (PACTE)